AFHRL-TP-90-6

AIR FORCE

HUMAN RESOURCES

LABORATORY

TOOL FOR STUDYING THE EFFECTS OF RANGE RESTRICTION IN CORRELATION COEFFICIENT ESTIMATION

Douglas E. Jackson

Eastern New Mexico University
Portales, New Mexico 88130

Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

July 1990

Interim Technical Paper for Period June 1988 - January 1990

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1990	Interim - June 1988 - January 1990

4. TITLE AND SUBTITLE
Tool for Studying the Effects of Range Restriction in Correlation Coefficient Estimation

5. FUNDING NUMBERS
PE - 62703F
PR - 7719
TA - 18
WU - 46

6. AUTHOR(S)
Douglas E. Jackson
Malcolm J. Ree

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432-1894

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFHRL-TP-90-6

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

It frequently happens that one must try to estimate the correlation coefficient between two random variables X and Y in some population P using data taken from a population Q, where Q is a proper subset of P. For example X and Y might be performance scores, P the set of individuals trying to gain acceptance into the armed services, and Q the subset of P consisting of those accepted. If X or Y or both are not part of the screening tests used as the basis for selection, then for at least one of these scores we have no data outside Q. We can administer tests to the members of Q and hence obtain data which may be used to estimate ρ_{X^*,Y^*} (X* and Y* are X and Y restricted to Q). Now suppose that Y is a criterion variable and we wish to measure the value of X as a means of selecting individuals who will have high Y scores. Obviously we want to know $\rho_{X,Y}$ and not ρ_{X^*,Y^*}. This paper has two purposes. The first is to present the equations involved in such a way that the problem becomes more intuitively understandable. The second is to describe a Monte-Carlo program written to simulate repeated sampling from Q. This program displays the sampling distribution of the traditional estimator for ρ_{X^*,Y^*} and of a proposed statistic for estimating $\rho_{X,Y}$. This proposed statistic is sometimes called the Pearson correction formula for range restriction. Currently the program assumes that the joint distribution of all variables is multinormal.

14. SUBJECT TERMS			15. NUMBER OF PAGES
coefficient	psychometrics	test theory	20
correlation	range restriction		16. PRICE CODE
estimation	simulation		

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 90-6 July 1990

TOOL FOR STUDYING THE EFFECTS OF RANGE RESTRICTION IN CORRELATION COEFFICIENT ESTIMATION

Douglas E. Jackson

Eastern New Mexico University
Portales, New Mexico 88130

Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

SUMMARY

It is sometimes necessary to estimate correlations in samples that have been range restricted due to selection. These correlations are often diminished when compared to their population values. A correction for this circumstance is the subject of a proof which is discussed in the context of a computer-aided simulation procedure to study the nature and behavior of the correction.

PREFACE

The present effort was completed as part of Work Unit 77191846, Development and Validation of Enlisted Selection Methodologies. It provides advanced statistical support for manpower programs.

The authors are grateful to the Air Force Office of Scientific Research for support of the effort. Drs. Valentine and Curran are thanked for their encouragement and facilitation of the effort.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

OOL FOR STUDYING THE EFFECTS OF RANGE RESTRICTION ON CORRELATION COEFFICIENT ESTIMATION

I. INTRODUCTION

Let X and Y be two random variables defined on a population P. Let Q be a sub-population of P and suppose that we have a random sample selected from Q. $X_1,...,X_n$ and $Y_1,...,Y_n$ will denote the X and Y data collected from this sample. The traditional statistic for estimating the correlation between X* and Y* [ρ_{X^*,Y^*}] is

$$r = \frac{\sum(X_I - \bar{X})\ (Y_I - \bar{Y})}{n - 1}.$$

However, this may not be a very good estimate of $\rho_{X,Y}$. The need to know $\rho_{X,Y}$ when you have only a random sample from Q is a problem that occurs quite naturally and it has been investigated for some time. The most widely used method to deal with it has been to use a correction formula first developed by Pearson (1903), and then extended by Lawley (1943). This formula applies when certain assumptions are satisfied. These assumptions are basically the classical linear regression model, and will be described in detail later. The formula applies only when ρ_{X^*,Y^*} is known exactly. It is not uncommon to take a formula that holds for population parameters and apply it instead to statistics used to estimate those population parameters. Unfortunately, this approach comes with no guarantees. It is not assured to provide an unbiased or even a very good estimation. Finding a mathematical description for the sampling distribution of the Pearson statistic appears to be very difficult. At least it has defied solution so far. Rather than seeking a mathematical solution, we decided instead to take a computational approach and write a Monte Carlo simulation program. The purpose of this program is to evaluate, under varying conditions, the accuracy of the traditional r statistic and of the Pearson statistic in estimating $\rho_{X,Y}$. It will also be useful for testing statistics that use correlation coefficients as inputs.

II. NOTATION AND OBJECTIVES

We will use the notation of Lord and Novick (1968). Assume that the members of an organization were admitted to the organization by virtue of having passed a battery of tests. These members are called the selected group or the restricted population and will be denoted by Q. These members plus those that were denied entry constitute the applicant group or the unrestricted population and will be denoted P. The tests that were used as a basis for selection are viewed as random variables on P and are called the explicit selection variables. Any other tests that are given to the members of the selected group Q are called incidental selection variables. All random variables are assumed to be defined on P. If X is a random variable on P, then the restriction of X to the selected group Q will be represented by the notation X*.

Our objective is to study the sampling distribution of two statistics. The first is the standard sample correlation coefficient, r, which is calculated using a random sample from Q. The second is the Pearson correction formula for range restriction. It is calculated using the sample covariance matrix for the explicit selection variables based on data from the applicant group P, plus the sample covariance matrix for all variables based on the selected group Q.

In this study it is assumed that the most general type of selection criterion is

$$\iota \leq c_1X_1 + \ldots + c_{nve}X_{nve} \leq h,$$

where *h* may be infinity, I may be negative infinity, and nve is the number of explicit selection variables.

III. CORRECTION IN THE TWO-VARIABLE CASE

The correction formula for range restriction is usually referred to as the Pearson correction formula. However, it was Lawley (1943) who established the minimum assumptions necessary for its application. In order to understand Lawley's theorem, it is necessary to look at a couple of special cases in this and the next section. The proof of the general theorem is by generating functions and is not a very instructive proof. The proof of the present special case, however, is instructive and an outline of this proof will be given.

Let X be the only explicit selection variable and let Y be the only incidental selection variable. Hence we have X and Y defined on P, and X* and Y* defined on Q; and the members of Q are selected on the basis of their X scores.

Assumption 1. (Linearity) The true regression function of Y on X is linear. In other words,

$$Y = a + bX + E,$$

where a and b are constants, E is a random variable, and the expected value of E given x is zero, for all x.

Note: It is not necessary to assume that X and E are independent. Linear regression is enough to imply that cov (X, E) = 0, which is needed for the proof of theorem 1. The proof that cov (X, E) = 0 follows directly from the definition of covariance and hence is omitted.

Assumption 2. (Homoscedasticity) The conditional variance of Y given x, does not depend on x. In other words, σ_E does not depend on x.

Note: Assumption 2 still does not imply that X and E are independent.

Theorem 1. Under assumptions 1 and 2

$$\rho^2_{X,Y} = \left[1 + \frac{s^2_{X^*}}{\sigma^2_X}\left(\frac{1}{\rho^2_{X^*,Y^*}} - 1\right)\right]^{-1}, \tag{1}$$

proof: Given that cov(X, E) = 0, it is a matter of simple algebraic manipulation and the relationship

$$\operatorname{cov}\left(\sum_i a_iX_i, \sum_j b_jY_j\right) = \sum_i\sum_j a_ib_j\operatorname{cov}(X_i, Y_j)$$

to show that

$$b = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X} \tag{2}$$

$$\sigma^2 E = \sigma^2 Y (1 - \rho^2 X,Y). \quad (3)$$

But now assumption 1 and Equation 2 imply that

$$\rho_{X,Y}\frac{\sigma_Y}{\sigma_X} = \rho_{X^*,Y^*}\frac{\sigma_{Y^*}}{\sigma_{X^*}}, \quad (4)$$

while assumption 2 and Equation 3 imply that

$$\sigma^2 Y (1 - \rho^2 X,Y) = \sigma^2 Y^* (1 - \rho^2 X^*,Y^*). \quad (5)$$

These two equations are exactly equivalent to the conclusion of the theorem. That is to say, you get the conclusion by solving for $\sigma^2 Y$ in Equation 4, putting that in Equation 5 and solving for $\rho_{X,Y}$.

It is important to understand that the conclusion depends exactly on linearity and homoscedasticity and the fact that Y is not explicitly restricted. No assumption of normality is needed. The population parameters that appear on the right side of the formula will, of course, not be known and so the statistic based on this theorem becomes

$$\left[1 + \frac{S^2_{X\cdot}}{S^2_X}\left(\frac{1}{r^2_{X\cdot,Y\cdot}} - 1\right)\right]^{-1},$$

which is the Pearson statistic for two variables. The sampling distribution of this statistic does depend on the joint distribution of X and Y. The simulation program described later assumes that this distribution is the bivariate normal. From looking at a few examples using that program it appears that the corrected statistic is always a slight underestimate. In the cases examined, this downward bias seems to be so small that it could easily be ignored.

Notice that if $\sigma^2 X^* < \sigma^2 X$, as it will be for the type of restrictions we are considering, then $\rho^2 X^*,Y^* < \rho^2 X,Y$. So if $r^2 X^*,Y^*$ is used instead of the correction formula, this gives an estimate for a parameter that is smaller than $\rho^2 X,U$. This last statement is not true when there are more than two variables. This will be discussed in the next section.

The effect of range restriction on population parameters in the two-variable case is easily visualized. Think of a correlation coefficient as a measure of how well we can perform the following task. Administer test X to two randomly chosen individuals and predict which of these individuals would score highest on test Y. There are three characteristics of the joint density function between X and Y that determine how well we can predict. The first is the slope of the regression line. This does not change with a restriction on X. The fact that it does not change is reflected by Equation 4, which is part of the proof of theorem 1. It is obvious that a greater slope leads to greater chance of successfully predicting which individual will have the larger Y score. The second is σE. This does not change with a restriction in X. The fact that it does not change is reflected in Equation 5, which is the other significant equation in the proof of theorem 1. It is clear that a smaller σE leads to a greater chance of picking the correct individual. The way this is reflected in the equation

$$\rho_{X,Y} = b\frac{\sigma_X}{\sigma_Y}$$

is that if σ_E is made smaller without changing σ_X or b, then σ_Y will become smaller. The third factor is the variance of X. It is clear that we have a better chance of choosing the correct individual if the X values of randomly chosen people are spread out rather than being packed together. This is the factor that is depressed as a result of selection. As already mentioned, for the type of selection in common use the variance of X* is always smaller than the variance of X. Hence in the two-variable case, selection always causes underestimation of the correlation coefficient if the correction formula is not used.

One of the benefits of presenting a proof of theorem 1 and then discussing how the three factors affect correlation coefficients is that one can see how correlation estimation depends on linearity and homoscedasticity. If one or both of these conditions fail drastically, then it will be very difficult to get a decent estimate. A few comments about this problem will be made at the end of this paper.

IV. THREE VARIABLES WITH ONE EXPLICITLY RESTRICTED

Let X be the explicit selection variable and let Y and Z be incidental selection variables. Y is the criterion or dependent variable and X and Z are the independent or predictor variables.

Assumption 1. The true regressions of Z on X, and Y on X, are linear.

Assumption 2. The variance of Y given X, the variance of Z given X, and the covariance of Z and Y given X, do not depend on X.

Theorem 2: Under assumptions 1 and 2

$$\rho_{Y\,Z} = \frac{\rho_{Y\cdot Z\cdot} + \rho_{X\cdot Y}\cdot\rho_{X\cdot Z}\cdot\left(\frac{\sigma_X^2}{\sigma_{X\cdot}^2} - 1\right)}{\left[1 + \rho_{X\cdot Y}^2\cdot\left(\frac{\sigma_X^2}{\sigma_{X\cdot}^2} - 1\right)\right]^{1/2}\left[1 + \rho_{X\cdot Z}^2\cdot\left(\frac{\sigma_X^2}{\sigma_{X\cdot}^2} - 1\right)\right]^{1/2}}$$

This correction formula is slightly more complex and it permits the construction of examples where the correlation in the restricted population is larger than the correlation in the unrestricted population. Levin (1972) refers to these cases as "Pseudo-Paradoxical." The terminology probably stems from the fact that it is widely assumed in the literature that restriction always causes an underestimate. This would certainly be expected on the basis of our discussion in the two-variable case. Notice that with the formula appearing in theorem 2 and a little algebra it is easy to characterize these "Pseudo-Paradoxical" situations in the three-variable case. In these cases the uncorrected estimation is an overestimation. Taking examples from Levin (1972), the correction procedure was applied and the simulation showed each time that the corrected value was very good. It seemed that the corrected estimate was slightly low in each case but the estimate was so close that this low estimation might not be a real effect. In any case the bias appears to be so slight that it is not practically significant. The interesting fact is that the correction statistic based on theorem 2 works well in these cases, at least when the joint distribution of the three variables is multinormal.

V. THE GENERAL CASE

Let X be the p-element vector of explicit selection variables, and Y the n-p element vector of incidental selection variables on the applicant group. Y will contain one criterion variable and several predictor variables. Then X* and Y* represent the explicit and incidental selection variables on the selected group. Let

$$V = \begin{bmatrix} V_{p,p} & V_{p,n-p} \\ V_{n-p,p} & V_{n-p,n-p} \end{bmatrix}$$

represent the variance-covariance matrix for X*, Y*. The first p rows and columns refer to the components of X*. So $V_{p,p}$ is the variance-covariance matrix of X*, $V_{n-p,n-p}$ is the variance-covariance matrix for Y*, V_{pn-p} gives the covariances between X* and Y*, and $V_{n-p,p}$ is the transpose of $V_{p,n-p}$. In this discussion, V refers to selected data and W refers to applicant data. In our application, V will be the estimate of the variance-covariance of all tests and it is based on selected data. The restricted population consists of those who were accepted into the organization so we have data on all tests for these people. Let

$$W = \begin{bmatrix} W_{p,p} & W_{p,n-p} \\ W_{n-p,p} & W_{n-p,n-p} \end{bmatrix}$$

be the matrix of variance-covariance for the unselected data. We will estimate $W_{p,p}$ from the data since we have data for the explicit selection variables on all applicants. The $W_{p,n-p}$, $W_{n-p,p}$, and $W_{n-p,n-p}$ are the matrices that we wish to know and will be given to us by the theorem. $W_{n-p,p}$ is, of course, the transpose of $W_{p,n-p}$; so, we will just give an expression for $W_{p,n-p}$ when we state the theorem. The following statement of the theorem is taken from Birnbaum, Paulson, and Andrews (1950).

Assumption 1: (Linearity) For each j the true regression of Y_j on X is linear.

Assumption 2: (Homoscedasticity) The conditional variance-covariance matrix of Y given X does not depend on X.

Theorem 3: Under assumptions 1 and 2

$$W_{p,n-p} = W_{p,p} V_{p,p}^{-1} V_{p,n-p} \quad \text{and}$$

$$W_{n-p,n-p} = V_{n-p,n-p} - V_{n-p,p}(V_{p,p}^{-1} - V_{p,p}^{-1} W_{p,p} V_{p,p}^{-1}) V_{p,n-p}$$

Lawley proved this theorem in 1943 using moment-generating functions. Both of the earlier theorems (1 and 2) are just special cases of theorem 3. With some algebraic manipulations the reader can verify this by writing out the entries of the matrix and comparing them with the formulas in the earlier theorems. Remember that the matrices of Lawley's theorem are variance-covariance matrices and so should be converted to correlation coefficients for the purposes of comparison.

Notice that the theorem says nothing about the types of restrictions that are allowed. Restrictions of any type on the X variables will preserve the linearity and homoscedasticity. However, if there are explicit restriction variables that are not known and hence not included in the equations of theorem 3, then the accuracy of the corrected statistic suffers. The conditions specified in Lawley's theorem are not met in this case.

VI. AN EXAMPLE

The following data are taken from Air Force performance measurement research records. They are the scores (Maier & Sims, 1986) on the 10 subtests of the enlistment qualification battery (variables 1-10) and a performance test (variable 11) (Green & Wing, 1988). Table 1 shows the correlations of these variables as observed in a sample from the restricted population. Hence they are not corrected for range restriction.

Table 1. Restricted Data

1.000										
0.143	1.000									
0.568	0.216	1.000								
0.381	0.244	0.537	1.000							
0.011	0.225	0.162	0.185	1.000						
0.112	0.251	0.200	0.157	0.583	1.000					
0.130	0.235	0.223	0.254	-.078	-.042	1.000				
0.371	0.381	0.271	0.192	0.244	0.320	-.125	1.000			
0.342	0.172	0.296	0.406	-.057	0.080	0.475	0.220	1.000		
0.308	-.030	0.161	0.018	-.158	-.039	0.451	-.018	0.283	1.000	
0.141	0.393	0.035	0.118	0.110	0.236	0.159	0.298	0.250	0.077	1.000

Table 2 shows the correlations of the first 10 variables (explicitly restricted variables) as calculated using a sample from the unrestricted population. Pearson correction will not change these values. They are the best estimates for the correlation coefficients between the explicitly restricted variables. Notice that some estimates have changed from slightly negative to significantly positive.

Table 2. Unrestricted Data

1.000									
0.722	1.000								
0.801	0.708	1.000							
0.689	0.672	0.803	1.000						
0.524	0.627	0.617	0.608	1.000					
0.452	0.515	0.550	0.561	0.701	1.000				
0.637	0.533	0.529	0.423	0.306	0.225	1.000			
0.695	0.827	0.670	0.637	0.617	0.520	0.415	1.000		
0.695	0.684	0.593	0.521	0.408	0.336	0.741	0.600	1.000	
0.760	0.658	0.684	0.573	0.421	0.342	0.745	0.585	0.743	1.000

Table 3 shows the correlations presented in Table 1 after the correction procedure has been applied. The last row of correlations of the subtests with variable 11 have now been corrected for range restriction. It is seen that these correlations have changed considerably in the process of being corrected for range restriction. These corrected values are the best available estimates for these correlation coefficients.

Table 3. Corrected Data

1.000										
0.722	1.000									
0.801	0.708	1.000								
0.689	0.672	0.803	1.000							
0.524	0.627	0.617	0.608	1.000						
0.452	0.515	0.550	0.561	0.701	1.000					
0.637	0.533	0.529	0.423	0.306	0.225	1.000				
0.695	0.827	0.670	0.637	0.617	0.520	0.415	1.000			
0.695	0.684	0.593	0.521	0.408	0.336	0.741	0.600	1.000		
0.760	0.658	0.684	0.573	0.421	0.342	0.745	0.585	0.743	1.000	
0.596	0.749	0.487	0.489	0.465	0.433	0.503	0.680	0.640	0.570	1.000

VII. GENERAL DESCRIPTION OF THE SIMULATION PROCESS

The program was written in PASCAL and is currently running on an IBM-compatible microcomputer. The joint distribution of all of the random variables is assumed to be multinormal in the unrestricted population. The inputs to the program are listed here for reference and they will be explained later as we discuss the program.

The number of variables [nv] and their names [vname]
Unrestricted population mean and std-dev [mu, sig]
The correlation coefficients in the unrestricted population [rho]
The number of explicitly selected variables [nve] the first nve entered
The number of restrictions [nr]
The coefficients of the explicitly selected variables [ncoeff]
Cutoff value for each restriction [cutoff]
Size of the unrestricted population [nwp]
Size of the restricted population [nvp]
The number of times the experiment will be repeated [reps]
The two variables of interest in the list of variables [int1,int2]

Figure 1 below is an example of a file describing the input to a run. The first line says that there are 3 variables in this case. The next three lines give the names, means, and standard deviations of the three variables. In this case they each have mean 0.0 and standard deviation 1.0. The next three lines give the correlation matrix for the three variables. So the coefficient for (x,y) is 0.86, for (x,z) it is 0.0, and for (y,z) it is 0.43. The next line gives the number of explicit selection variables. There is 1 in this case and so X is the only explicit selection variables. Then it is specified that there is only 1 restriction (selection) and that the restriction is $X \geq 0.0$.

The selected group will consist of those persons getting a score of zero or greater on the X test. The second to last line says that the variables of interest are 2 and 3 (i.e., Y and Z). Data and a histogram of the distribution will be given for the uncorrected r between X and Z and the same information is given for the Pearson correction statistic. The program calculates the Pearson correction statistic using theorem 3 from the last section. The line will be explained after the following discussion.

```
3
x   0.0     1.0
y   0.0     1.0
z   0.0     1.0
      1.00    0.86    0.00
      0.86    1.00    0.43
      0.00    0.43    1.00
1     # of explicitly restricted variables
1     number of restrictions
1.0   0.0
2 3   variables of interest
1 50 100
```

Figure 1. **An Input File.**

Creating a multinormal observation is equivalent to simulating one individual. In the above case this means getting three values,--one for each of the three test scores X, Y, and Z. Each multinormal observation is part of the applicant group and is also a member of the selected group if the scores satisfy all of the restrictions. For the present case this means that the score on the X test must be at least zero. One experiment is simulated by generating observations until two conditions are satisfied. There must be at least nwp observations in the applicant group and there must be at least nvp observations in the selected group. For most cases we set nwp = 1 and then the only restriction is that we have at least nvp observations in the selected group. One run of the program consists of simulating reps experiments. The last line of a file which describes a run gives, nwp, nvp and reps in that order. In Figure 1, nwp = 1, nvp = 50, and reps = 100.

When program corr begins, it will ask if the user wants to enter the data necessary to describe a run or to give the name of a file which contains the data in the expected format. The file in Figure 1 is called test4 and so we can just give that name to corr and the run is specified by the input parameters in Figure 1. The reason that test4 is in the expected format is that corr wrote the file on a previous run. It was written when corr executed and it was specified that data would be entered from the keyboard and that these data were to be saved in a file named test4. Now if one were familiar with PASCAL read statements, they could use a text editor to change some of the parameters and use test4 for another run. After corr executes, the data necessary to produce the histograms of the corrected and the uncorrected statistics are in two internal files and one must run program plot which will read these internal files and display these data on the printer.

For each experiment corr calculates each of the following quantities.

```
b0 and b1 = the estimates of the regression parameters
statu = the uncorrected estimate of the correlation coefficient
statc = the corrected estimate of the correlation coefficient calculated
          with the equations of theorem 3
```

Hence corr will generate reps copies of each of these parameters. In each case the two implied variables are int1 and int2, and the regression parameters are for int2 on int1. In the case of b0 and b1, the only values retained are the totals so that after the reps experiments have been generated, the mean values of these parameters may be calculated. In the case

...tatc, each observed value is retained and written to the files pltu.dat and pltc.dat, ...spectively. As mentioned earlier, the user can run plot to have all these results displayed.

VIII. PROGRAM METHODOLGY

One can see that the correction procedure, as specified in theorem 3, requires taking the inverse of a matrix. This is accomplished with the Gauss-Jordan matrix inversion algorithm in unit matops. This unit also contains algorithms to multiply and to subtract matrices.

Unit normgen includes all of the routines necessary to generate a multinormal observation with the correlations specified in the input file. Suppose that there are nv variables. The first step is to generate nv independent standard normal observations. This is accomplished by repeated calls to algorithm p in Knuth (1969). The desired multinormal distribution results from taking a linear transformation of these independent standard normal observations. This transformation is obtained by multiplying the independent observations and the matrix A which is defined to be that unique matrix which is upper triangular and satisfies $AA^T = C$. In this last equation, A^T refers to the transpose of A, and C is the variance-covariance matrix of all variables in the unrestricted population. For a complete discussion of this procedure, consult Shreider (1966) or Johnson (1987). The matrix A is calculated by the recursive procedure solve called by transpar in unit normpar.

IX. RECOMMENDATIONS

Much time has been spent in writing the program; hence, most of these recommendations have to do with proposed applications of the tool. However, based on a limited amount of experimentation, a few observations seem appropriate.

The correction statistic seems to work well under the conditions of the theorem. It seems to have a downward bias but, for the cases we considered, it was always preferable to the uncorrected statistic. As can be seen from the proof of theorem 1, neither the corrected nor the uncorrected statistic will be accurate if the joint distribution of all variables fails to satisfy the linearity condition or the homoscedasticity condition. After fully understanding the theorem, and a little experimentation with the simulation program, it seems likely that the best strategy is to always use the Pearson correction statistic instead of the uncorrected statistic.

There are a number of studies that could be pursued with the use of the simulation program. A plot of the sampling distribution of the Fisher Z-transformation of the corrected statistic looked approximately normal, as might be expected. The Z-transformation could form the basis for a procedure that could be used to construct confidence intervals for the true correlation coefficient based on the Pearson statistic. It might be instructive to modify the program slightly so as to allow the joint distribution of all variables to be specified in the input. This would allow one to test the confidence intervals procedure using actual data instead of stochastically generated multinormal data.

It would be useful to know how much accuracy is lost in the corrected statistic when one or more explicit selection variables have been omitted from the model. Based on a few experiments, the accuracy of the corrected statistic is diminished by the omission of explicit selection variables. It would be of value to know the magnitude of this effect. This is important since some people still use the two- or three-variable formulas even when there is more than one explicit selection variable. The simulation program is ideally suited to answer this question.

REFERENCES

Birnbaum, Z.W., Paulson, E., & Andrews, F.C. (1950). On the effects of selection performed on some coordinates of a multi-dimensional population. *Psychometrika*, 15(2), 191-204.

Green, B.F., & Wing, H. (Eds.), (1988). *Analysis of job performance measurement data*. Report of a workshop, Washington, DC: National Academy Press.

Johnson, M.E. (1987). *Multivariate statistical simulation* (p.49). New York: John Wiley and Sons.

Knuth, D. (1969). *The art of computer programming: Vol.2. Seminumerical algorithms* (p.104). Menlo Park, CA: Addison-Wesley.

Lawley, D.N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edingburg*. Sect. A. (Math & Psys. Sec.), 62, Part I, 28-30.

Levin, J. (1972). The occurence of an increase in correlation by restriction of range. *Psychometrika*, 37(1), 93-97.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Maier, M.H., & Sims W.H. (1986). *The ASVAB score scale: 1980 and World War II (CNR 116)*. Arlington, VA: Center for Naval Analysis.

Pearson, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Phil. Transactions of the Royal Society A*, 200, 1-66.

Shreider, Y.A. (1966). *The Monte Carlo method* (p. 328). Elmsford, NY: PerGamon Press.